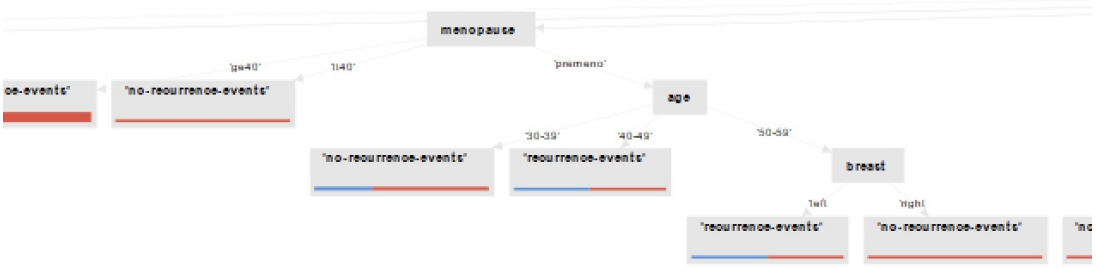
Homework 2

DSTBD_Quaderno2-ITA.pdf

1. Generare un albero di decisione con l’algoritmo Decision Tree usando l’intero dataset per il training, settando il minimal gain a 0.01 e mantenendo la configurazione di default per gli altri parametri.

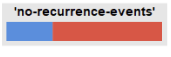
   a. Quale attributo è considerato dall’algoritmo il più selettivo al fine di predire la classe di un nuovo dato di test?
   L’attributo più selettivo per la predizione è “node-caps”, in quanto è la radice dell’albero.

   b. Qual è l’altezza dell’albero di decisione generato?
   Per altezza di un albero si intende il numero massimo di archi che separano la radice da una foglia.
   Nel nostro caso l’altezza è 5, come si può vedere nel seguente caso:
   node-caps = 'no’ → irradiat = 'no’ → tumor-size = '15-19’ → menopause = 'premeno’ → age = '50-59’ → breast = 'right': 'no-recurrence-events'
   {'recurrence-events'=0, 'no-recurrence-events'=2}



c. Trovare un esempio di partizionamento puro all'interno dell'albero di decisione generato e riportare uno screenshot che mostri l'esempio trovato. Un partizionamento, ossia la divisione dei dati in base al valore di un attributo, viene detto puro quando tutti i dati appartengo ad uno solo dei valori dell'attributo. L'analisi del partizionamento viene utilizzata per scegliere i migliori attributi su cui fare lo split, il partizionamento puro, quando fatto su un attributo che generalizza, è ottimale.

2. Analizzare l'impatto del minimal gain (considerando il gain ratio come criterio di splitting) e del maximal depth sulle caratteristiche dell'albero di decisione generato dall'intero dataset (mantenendo la configurazione di default per gli altri parametri di configurazione).
   Riportare almeno 5 screenshot differenti che mostrino gli alberi di decisione (o porzioni di essi) generati con differenti configurazioni.
   Il minimal gain è il parametro che permette di scegliere se fare un ulteriore split dell'albero. Il minimal gain rappresenta la soglia minima oltre cui si può splittare l'albero.
   In un albero la profondità di un nodo è la lunghezza del percorso dalla radice al nodo stesso, ossia il numero di archi tra la radice e il nodo, mentre l'altezza di un albero è la profondità massima.
   Valori elevati di minimal gain producono un numero limitato di partizionamenti e, di conseguenza, alberi di decisione più piccoli. Valori troppo elevati di minimal gain (ad es., 0.9) impediscono completamente lo split dei valori degli attributi e quindi l'albero risultante conterrà un singolo nodo.

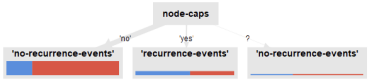


Max depth: 5, minimal-gain: 0.5

Max-depth: 4, minimal gain: 0.05

Max-depth: 4, minimal gain: 0.01

Max-depth: 10, minimal gain: 0.001



Max-depth: 2, minimal gain: 0.02

3. Applicando un 10-fold Stratified Cross-Validation, qual è l'effetto del minimal gain e del maximal depth sull'accuratezza media ottenuta da Decision Tree?
   Riportare almeno 5 screenshot che mostrino le matrici di confusione ottenute usando diverse configurazioni per i parametri sopra citati (considerare almeno le 5 configurazioni usate per rispondere alla domanda 2). Mantenere la configurazione di default per tutti gli altri parametri.

accuracy: 70.30% +/- 1.43% (micro average: 70.28%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	0	0	0.00%
pred. 'no-recurrence-events'	85	201	70.28%
class recall	0.00%	100.00%	

Max depth: 5, minimal-gain: 0.5

accuracy: 71.33% +/- 6.58% (micro average: 71.33%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	26	23	53.06%
pred. 'no-recurrence-events'	59	178	75.11%
class recall	30.59%	88.56%	

Max-depth: 4, minimal gain: 0.05

accuracy: 71.00% +/- 6.95% (micro average: 70.98%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	31	29	51.67%
pred. 'no-recurrence-events'	54	172	76.11%
class recall	36.47%	85.57%	

Max-depth: 4, minimal gain: 0.01

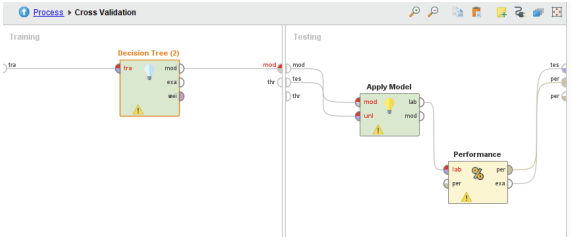
accuracy: 67.48% +/- 6.59% (micro average: 67.48%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	37	45	45.12%
pred. 'no-recurrence-events'	48	156	76.47%
class recall	43.53%	77.61%	

Max-depth: 10, minimal gain: 0.001

accuracy: 68.90% +/- 6.96% (micro average: 68.88%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	28	32	46.67%
pred. 'no-recurrence-events'	57	169	74.78%
class recall	32.94%	84.08%	



4. Considerando il classificatore K-Nearest Neighbor (K-NN) e applicando un 10-fold Stratified CrossValidation, qual è l'effetto del parametro K sull'accuratezza media del classificatore?
   Riportare almeno 5 sceenshot che mostrino le matrici di confusione ottenute usando diversi valori di K. Applicare un 10-fold Stratified Cross-Validation con il classificatore Naïve Bayes. K-NN ottiene mediamente prestazioni superiori o

inferiori a Naïve Bayes classifier sul dataset
analizzato? Riportare uno screenshot che mostri la matrice di confusione ottenuta con Naive Bayes sul dataset analizzato.

accuracy: 73.77% +/- 5.98% (micro average: 73.78%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	26	16	61.90%
pred. 'no-recurrence-events'	59	185	75.82%
class recall	30.59%	92.04%	

k=5

accuracy: 65.73% +/- 8.62% (micro average: 65.73%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	39	52	42.86%
pred. 'no-recurrence-events'	46	149	76.41%
class recall	45.88%	74.13%	

k=2

accuracy: 75.20% +/- 5.43% (micro average: 75.17%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	25	11	69.44%
pred. 'no-recurrence-events'	60	190	76.00%
class recall	29.41%	94.53%	

k=10

accuracy: 73.79% +/- 5.61% (micro average: 73.78%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	17	7	70.83%
pred. 'no-recurrence-events'	68	194	74.05%
class recall	20.00%	96.52%	

k=20

accuracy: 74.51% +/- 5.02% (micro average: 74.48%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	12	66.67%
pred. 'no-recurrence-events'	61	189	75.60%
class recall	28.24%	94.03%	

k=8

5. Analizzare la matrice di correlazione per valutare la correlazione tra coppie di attributi del dataset. Riportare uno screenshot che mostri la matrice di correlazione ottenuta. Alla luce dei risultati ottenuti, l'ipotesi d'indipendenza Naïve risulta valida per il dataset Breast? Qual è la coppia di attributi maggiormente correlati?
   La matrice di correlazione riporta la correlazione simmetrica tra gli attributi, che si trovano nell'asse x e nell'asse y, motivo per cui tutti gli elementi della diagonale della matrice valgono sempre 1. L'ipotesi Naïve è realistica quando non sussistono correlazioni significative tra le coppie di attributi.

Attribut...	age	menopa...	tumor-s...	inv-nodes	node-ca...	deg-mal...	breast	breast-...	irradiat
age	1	?	?	?	?	?	?	?	?
menopa...	?	1	?	?	?	?	?	?	?
tumor-size	?	?	1	?	?	?	?	?	?
inv-nodes	?	?	?	1	?	?	?	?	?
node-caps	?	?	?	?	1	?	?	?	?
deg-malig	?	?	?	?	?	1	?	?	?
breast	?	?	?	?	?	?	1	?	-0.019
breast-q...	?	?	?	?	?	?	?	1	?
irradiat	?	?	?	?	?	?	-0.019	?	1